Test a Witness's Memory of a Suspect Only Once

John T. Wixted, Gary L. Wells, Elizabeth F. Loftus, & Brandon L. Garrett

University of California, San Diego

Accepted for publication: *Psychological Science in the Public Interest*

Author Note

John T. Wixted, Department of Psychology, University of California, San Diego.

Correspondence concerning this article should be addressed to John Wixted (jwixted@ucsd.edu).

Abstract

Eyewitness misidentifications are almost always made with high confidence in the courtroom. The courtroom is where eyewitnesses make their *last* identification of defendants suspected of (and charged with) committing a crime. But what did those same eyewitnesses do on the *first* identification test conducted early in a police investigation? Despite testifying with high confidence in court, many eyewitnesses also testified that they had initially identified the suspect with low confidence or failed to identify the suspect at all. Presenting a lineup leaves the eyewitness with a memory trace of the faces in the lineup, including that of the suspect. As a result, the memory signal generated by the face of that suspect will be stronger on a later test involving the same witness, even if the suspect is innocent. In that sense, testing memory contaminates memory. These considerations underscore the importance of a newly proposed recommendation for conducting eyewitness identifications: *Avoid repeated identification procedures with the same witness and suspect*. This recommendation applies not only to additional tests conducted by police investigators but also to the final test conducted in the courtroom, in front of the judge and jury.

Key words: Eyewitness Identification; Wrongful Convictions; Malleability of Memory

Test a Witness's Memory of a Suspect Only Once

"No man ever steps in the same river twice, for it's not the same river and he's not the same man" -- Heraclitus

In a court of law, a credible eyewitness who confidently identifies a defendant as the culprit of a crime is often thought to provide direct and powerful evidence of guilt. Indeed, judges have traditionally characterized a courtroom identification as having an "independent" and direct "source" in the witness's memory. Although underappreciated in the legal system, despite being almost universally understood by experimental psychologists, an eyewitness identification in court does *not* provide direct evidence of guilt. Nor is it independently sourced in the witness's memory. Instead, by the time of trial, an eyewitness's memory has almost invariably been contaminated by a variety of factors and is therefore highly error prone. As of today, 375 prisoners have been exonerated by DNA testing, 21 of whom were on death row, and it is now widely understood that eyewitness misidentifications contributed to ~70% of these wrongful convictions (Innocence Project, 2020).

Eyewitness misidentifications typically first occur during early stages of a police investigation (e.g., when a lineup is administered), long before trial. Decades of research have therefore focused on proper methods for conducting lineups in such a way as to minimize initial misidentifications. As the relevant research accumulated over the years, consensus science-based recommendations about proper eyewitness identification procedures have evolved accordingly. The most recent set of guidelines set forth by Wells, Kovera, Douglass, Brewer, Meissner, & Wixted (2020) include a new recommendation that is the focus of this article. Specifically, Recommendation #8 is as follows: *Avoid repeated identification procedures with the same witness and suspect*. In other words, test a witness's memory for a suspect only once.

Under the right conditions, the first eyewitness identification test can provide reliable information. According to a recent review of the literature, on an *initial* lineup identification test of *uncontaminated* memory conducted in accordance with current recommendations (i.e., when a pristine procedure is used), confidence can be a reliable indicator of accuracy (Wixted & Wells, 2017). That is, a high-confidence identification implies high accuracy, whereas a low-confidence identification implies low accuracy (the veritable definition of eyewitness reliability). How often pristine conditions prevail in the real world is unknown, but what is known is that, on the *first* test, eyewitness identification evidence is potentially reliable.

No later test provides more reliable information than the first test because memory is malleable (Davis & Loftus, 2018). That is, as with other forms of forensic evidence, memory can be contaminated. Critically, the best chance to test uncontaminated memory is the first test because the very act of testing memory can contaminate it (Steblay and Dysart, 2016). The importance of testing a witness's memory for a suspect only once is hard to overemphasize because the failure to abide by that simple rule might account for a large proportion of the wrongful convictions overturned by DNA evidence (Garrett, 2011). Later, we detail specific research findings and representative real-world cases supporting these claims.

Implementing this newly proposed reform should be simple and straightforward because it involves no special training beyond educating police investigators, prosecutors, and judges about its compelling science-based rationale. The purpose of this article is to do just that. We begin by tracing the growing awareness of the importance of the recommendation to avoid repeated testing by briefly reviewing how consensus science-based guidelines for conducting eyewitness identification procedures have evolved over the years.

The evolution of guidelines pertaining to eyewitness identification procedures

A team of scientists (sometimes working with law enforcement and legal practitioners) has been commissioned to draw up recommendations for conducting eyewitness identification procedures four times, beginning in 1998. The first guidelines were enumerated in a "white paper" (Wells, Small, Penrod, Malpass, Fulero, & Brimacombe, 1998) commissioned by the American Psychology-Law Society (APLS). That paper provided four recommendations for conducting police lineups consisting of one suspect and several physically similar fillers. Nowadays, the police typically use photo lineups instead of the live lineups that were once the norm (Police Executive Research Forum, 2013). The four recommendations in that white paper were as follows: (1) the lineup administrator should be blind to the identity of the suspect, (2) the eyewitnesses should be informed that the culprit may or may not be in the lineup, (3) the suspect should not stand out in the lineup (i.e., the lineup should be fair), and (4) a confidence statement should be obtained at the time an identification is made and prior to any feedback from the police. All of these recommendations remain in force today, but as of 1998, the importance of testing memory only once was not yet apparent.

One year later, the National Institute of Justice (NIJ) issued another set of science-based guidelines (Technical Working Group for Eyewitness Evidence, 1999). Whereas the 1998 white paper focused on lineups, per se, the NIJ guidelines were much broader, providing recommendations for creating mug books and composites, for interviewing eyewitnesses, for conducting showups (which involve only the suspect), and for conducting lineups. The lineup recommendations were similar to those in the 1998 white paper, though with added specificity on some issues (e.g., the recommendation that at least five fillers be included in a lineup). Still,

no mention was made was made about the special importance of testing a witness's memory for a suspect only once.

In 2013, a committee was appointed by the National Academy of Sciences to provide updated recommendations for eyewitness identification tests (National Research Council, 2014). Some of the new recommendations emphasized system-level issues such as training law enforcement officers in eyewitness identification procedures and conducting pretrial judicial inquiries into the reliability of the eyewitness evidence. With regard to eyewitness identification procedures per se, they reiterated some of the earlier recommendations and added others, such as the recommendation that the eyewitness identification procedure be videotaped. Critically, they also added a new recommendation that reflected increased awareness of the importance of the initial identification. Specifically, their recommendation #7 is as follows: "Make Juries Aware of Prior Identifications." In justifying this new recommendation, the committee wrote: "In-court confidence statements may also be less reliable than confidence judgments made at the time of an *initial* out-of-court identification; as memory fails and/or confidence grows disproportionately" (p. 110, emphasis added). They also noted that "Eyewitness testimony is a type of evidence where (as with forms of forensic trace evidence) contamination may occur pre-trial" (p. 109). Contamination is the crux of the issue.

The next major development occurred when the APLS commissioned Wells et al. (2020) to update the 1998 white paper in light of what has been learned since that time. There are now nine recommendations, including such new recommendations as conducting a pre-lineup interview (in part to warn the witnesses against attempting to identify the culprit on social media and elsewhere) and, as noted above, to avoid repeated identifications with the same witness and same suspect. The overarching reason to avoid repeated tests is that memory is malleable. The

essential problem is that on a second test, an individual can look familiar because of the exposure during the first test, even when it is not the right person. Next, we consider how the field came to appreciate that fact and how it leads to the conclusion that law enforcement should avoid testing a witness's memory for a suspect more than one time.

Memory is Malleable

Concerns about the malleability of memory can be traced back to at least Munsterberg (1908), but a scientific consensus about how easily memories can be modified—or even manufactured outright—did not begin to emerge until the mid-1970s. At that time, Loftus and Palmer (1974) and Loftus, Miller, and Burns (1978) reported the once surprising but now widely accepted finding that something as subtle the nature of a question posed to an eyewitness can influence what the witness later remembers. Subsequent studies showed that people can even be induced to falsely remember entire events that never happened, such as being lost in a shopping mall as a child (Loftus & Pickrell, 1995) or that they were attacked by a vicious animal (Porter, Yuille, & Lehman, 1999).

The examples summarized above pertain to memory tested by recall (i.e., recollecting details pertaining to a prior event), but eyewitness identification is a *recognition* memory test. As noted earlier, the malleability of memory has proven to be a particularly pernicious force on these tests, with many of the DNA exonerations involving eyewitnesses who incorrectly "recognized" the innocent suspect as the culprit. A striking example of memory contamination in the context of recognition memory was reported by Morgan, Southwick, Steffian, Hazlett, and Loftus (2013). They conducted a study of military personnel who were confined to a mock prisoner-of-war camp during survival school training. Each trainee experienced ~30 minutes of physically confrontational interrogation while alone in a room with an instructor. After the

interrogation, the trainee was left alone in an isolation cell. Later, a member of the research team entered the cell and asked questions about the interrogator ("Did your interrogator give you anything to eat?") while showing the participant a photograph of a Caucasian male (the "foil"), thereby falsely implying that he was the interrogator.

Next, memory for the interrogator was tested using a 9-person target-absent simultaneous photo lineup. The photo lineup contained a picture of the foil but not the actual interrogator (i.e., it was a target-absent lineup in which the foil's face had been differentially familiarized under highly suggestive conditions). Participants who had not been exposed to the foil's face following interrogation identified the foil as the interrogator 15% of the time. By contrast, participants who had been exposed to the foil's face identified the foil as the interrogator a remarkable 84% of the time.

As the example presented above clearly indicates, testing memory with suggestive or otherwise improper procedures contaminates memory. A natural assumption might be that testing memory under optimal conditions (i.e., in accordance with current recommendations) would not have a contaminating effect. This may very well be true when the memory test in question consists of interviewing a witness about their recollection of details about the crime (a recall test) using a proper procedure such as the Cognitive Interview (Fisher & Geiselman, 1992). When questioned properly, witnesses tend to recall accurate information. Recalling accurate information does not contaminate memory. Indeed, it can reinforce it by making the retained information more durable than it otherwise would be, a memory-enhancing phenomenon known as "the testing effect" (e.g., Roediger & Karpicke, 2006).

Unfortunately, the same is not true when memory is tested using a recognition procedure such as a lineup. Even when using a pristine lineup procedure that happens to involve an

innocent suspect, testing memory generally contaminates memory for that individual, thereby rendering any later recognition test prejudicial. Next, we consider some theoretical concepts derived from years of basic-science research to understand how and why that happens.

A Primer on the Theoretical Understanding of Recognition Memory

Several longstanding and influential theoretical considerations help to make sense of recognition memory: (1) encoding specificity, (2) similarity-based matching, (3) elaborative processing, (4) signal detection theory, and (5) the source-monitoring framework. These are all standard “textbook” ideas that inform our understanding of the intuitively simple but surprisingly complex act of recognizing a once-seen face. The theoretical issues discussed in the remainder of this section are outlined in Figure 1.

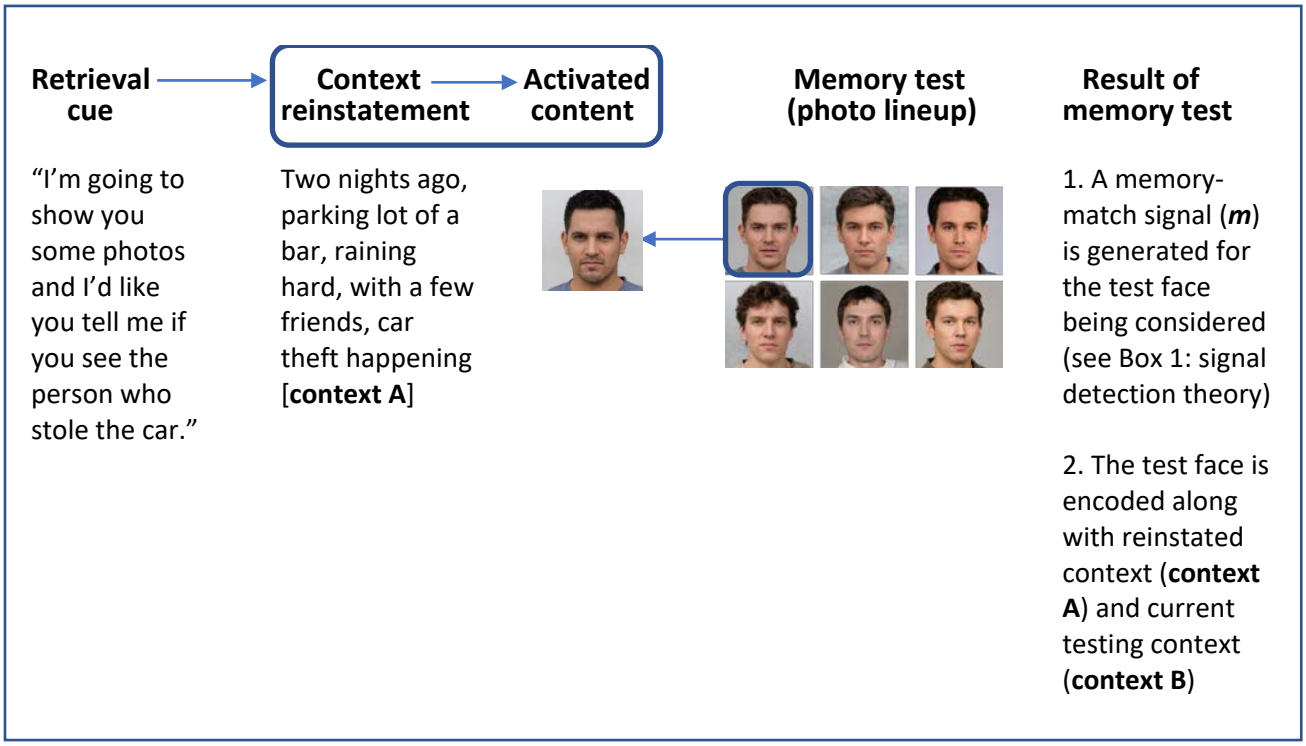


Figure 1. The specific reference to the witnessed crime by the investigating officer (left column) is a retrieval cue that reinstates the encoding context in the mind of the eyewitness (context A), which activates the relevant content—namely, the memory of the culprit. If multiple culprits were involved, their faces would also be activated, as would faces similar to the culprit(s) that might have been seen by the witness in other contexts. We omit those considerations for simplicity. Next, the lineup administrator presents the photo array to the witness, and each photo is compared to the activated content to make an identification decision. This figure illustrates that comparison process taking place for the top-left photo in the lineup. The comparison process yields a memory-match signal associated with the tested face (*m*) that is conceptualized in terms of signal detection theory (Box 1). After all the faces in the array have been compared to memory, there will be six memory-match signals, and the face associated with the strongest signal will be a candidate for being identified. Comparing a face to memory involves elaborative processing and so incidentally creates a distinctive memory of the tested face (a process illustrated in Figure 2), one that is encoded along with aspects of the reinstated context (A) and the testing context (context B).

Encoding specificity

Memory is generally understood to be cue-dependent (Tulving, 1983; Tulving & Thomson, 1973), which is to say that what you explicitly remember is determined by a *retrieval cue* that distinguishes the sought-after memory from the multitude of memories stored in one's brain. When memory is tested using a lineup, the retrieval cue consists of the specific question put to the witness. This is important because memories are differentially activated and thus accessible depending on the cues available at test (e.g., Godden & Baddeley 1975). The question posed to the witness is not (or should not be) "are any of these faces familiar?" Instead, the more direct question is: "do you see the person who committed the crime?" That retrieval cue will reinstate the context of the crime and activate the relevant content (namely, the face of the culprit) in the brain of the eyewitness, as illustrated in the leftmost columns of Figure 1.

Similarity-Based Matching

In the simplest and perhaps most common case, the activated content consists of only one face (the singular culprit). If, instead, multiple culprits were involved, all their faces would be activated. According to global matching models, beginning with Gillund and Shiffrin (1984), each recognition test item (e.g., each face in the lineup) is separately and individually compared against the activated faces (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). In the case of a single culprit, this process reduces to what one might already intuitively assume to be true: each face in the lineup is separately compared against the remembered face of the culprit. Figure 1 illustrates the comparison process for one face in the lineup, which yields a memory-match signal (***m***) for that face. This memory signal will be stronger the more similar the face in the lineup is to the witness's memory of the culprit.

Signal detection theory

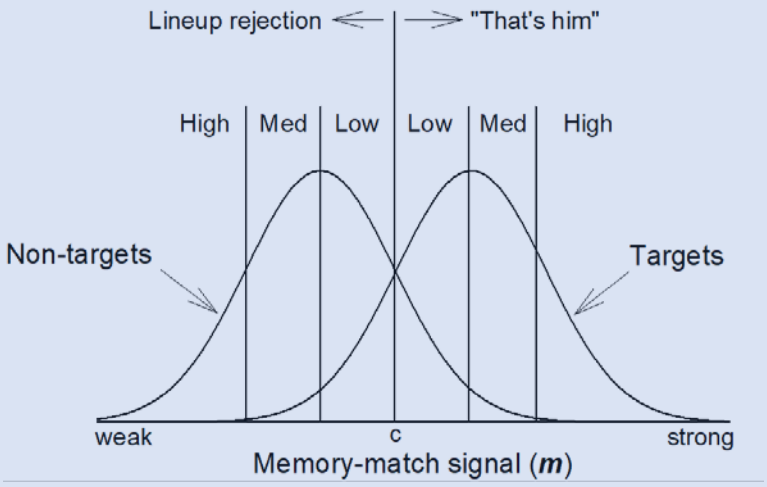
The memory-match signal (***m***) associated with a tested face is usually conceptualized in terms of signal detection theory. According to this theory, the memory signal is not all-or-none (match vs. no match) but is instead continuous because a face in the lineup can have any degree of similarity to the face of the culprit in memory. It seems natural to assume that, in a target-present lineup, the guilty suspect's face will be the most similar (generating the highest value of ***m***), whereas the fillers will be less similar (generating weaker values of ***m***). Similarly, in a fair target-absent lineup, no one will be very similar to the face of the culprit in memory, so they should all generate weak values of ***m***. These assumptions are sensible, but they omit an important consideration. According to signal detection theory, only *on average* is a guilty suspect expected to generate a stronger memory signal than an innocent suspect or a filler. The reason is that memory matching is an inherently noisy process (Box 1). Thus, occasionally (but not usually), a guilty suspect will generate a weak memory signal and an innocent suspect (or a filler) will generate a strong memory signal. A troubling implication is that, even under ideal conditions involving no memory contamination and pristine testing procedures, and even on the initial test, misidentifications will inevitably happen from time to time. Still, high-confidence misidentifications should be rare. However, for reasons explained next, misidentifications would be expected to increase if memory is tested a second time.

Box 1: Signal Detection Theory

Signal detection theory is a conceptual framework with origins dating back to the dawn of experimental psychology (Fechner, 1860; Kellen et al., in press; Green & Swets, 1966; Wixted, 2020). As applied to eyewitness identification, signal detection theory conceptualizes the memory-match signal (***m***) that is generated in the brain of an eyewitness when a face in a lineup (innocent or guilty) is compared to the face of the culprit stored in memory. The more similar the two faces are, the stronger the memory-match signal will be. On average, ***m*** will be strong when the face under consideration is the guilty suspect (the *target*) because that face matches memory of the culprit, but it will not always be strong. This means that we should think of ***m*** not as a constant but as a variable that has a range of values across many eyewitnesses who are considering the guilty suspect in a target-present lineup. Its value will be high on average, but it will have variance as well. Thus, in signal detection theory, we represent ***m*** for guilty suspects as a *distribution* of values with a relatively high mean.

When the face under consideration is an innocent suspect in a target-absent lineup or a filler in either type of lineup (*non-targets*), ***m*** will be weak, on average, because these faces will not usually be particularly similar to the memory of the culprit. However, it will not always be weak (e.g., the innocent suspect might be a lookalike). Thus, once again, across many lineups and eyewitnesses, it is useful to conceptualize ***m*** as a variable with a relatively low mean, not as a constant. Thus, there are two distributions (assumed to be Gaussian in form and with equal variance for convenience) with different means, one for guilty suspects and another for both innocent suspects and fillers. For measurement purposes, we can conceptualize the difference between the target and non-target means in standard deviation units (σ) and call that measure d' (a value estimated from data in a particular experiment).

After examining all the faces in the lineup, one face will have the maximum value of ***m*** ($\boldsymbol{m}_{MAX}$). If $\boldsymbol{m}_{MAX}$ is strong enough—that is, if it exceeds the witness's *decision criterion* (c) for making an ID—that face is identified Wixted, Vul, Mickes, & Wilson, 2018). If so, the stronger $\boldsymbol{m}_{MAX}$ is, the higher the witness's confidence in that ID will be (low, medium or high). If even $\boldsymbol{m}_{MAX}$ is not strong enough to exceed the witness's decision criterion, the lineup is rejected. The weaker $\boldsymbol{m}_{MAX}$ is, the higher the witness's confidence that the culprit is not present in the lineup. This model inherently predicts a strong confidence-accuracy relationship (Wixted, 2020), which is often observed in lab studies for suspect IDs (Wixted & Wells, 2017) but is less reliably observed for lineup rejections (e.g., Brewer & Wells, 2006).



Elaborative Processing

The comparison process between a particular face in the lineup and the activated content of the culprit's face in memory does more than simply yield a memory-match signal. It also creates a detailed memory record because of the face-processing that occurred during the identification procedure. In a typical lineup, the suspect and the fillers will be physically similar to each other. For example, to be included in the lineup, the face would ideally match the description of the perpetrator provided by the eyewitness (e.g., clean-shaven 20-year-old white male with short dark hair). Because of how lineups are designed, it will not suffice to perform a superficial scan of each face to make an identification decision, such as only taking notice of the shared features. Instead, each face in the lineup must be more thoroughly processed than that by attending to additional dimensions of the face (Figure 2).

Critically, the act of attending to additional facial dimensions means that the witness has processed some of the unique features that, in configuration, define how a face in the lineup differs from other faces in the population. In other words, by necessity, a face in a lineup is *elaboratively processed* to decide if this is the person who committed the crime. Such elaborative processing takes place whether the ultimate identification decision is "yes" or "no." Decades of research have established that the more elaboratively a stimulus is processed, the more likely it is to be later remembered (Craik & Tulving, 1975). Why? Craik (2002) put it this way: "…a richly elaborate trace will be more differentiated from other episodic records—this greater distinctiveness in turn will support more effective recollection in an analogous way to distinctive objects being more discriminable in the visual field" (pp. 306-307).

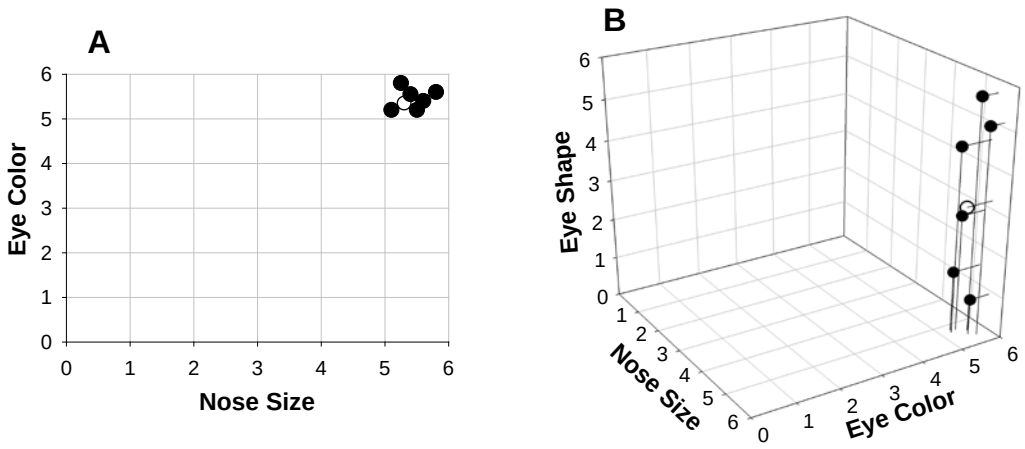


Figure 2. Multidimensional "face space" (Valentine, 1991; Valentine, Lewis, & Hills, 2016). Although face perception and memory are widely thought to involve both holistic and feature-based processing (Abudarham, Shkillera, & Yovel, 2019; McKone & Yovel, 2009; Chua, Richler, & Gauthier, 2015; Tanaka & Farah, 1993; Tanaka, & Simonyi, 2016), low-level perceptual features are used as perceived facial dimensions here for illustrative purposes. The dimensions could just as easily reflect more global properties of any level of abstractness (e.g., masculinity, attractiveness, perceived trustworthiness, etc.).

Panel A: In perception, representation of six members of a target-present lineup (filled circles) and, in memory, representation of the culprit (open circle) on two facial feature dimensions, Nose Size and Eye Color. In this hypothetical example, all values fall between 5 and 6 on both dimensions because the witness described the culprit as having a large nose (0 = very small to 6 = very large) and dark brown eyes (0 = very light blue to 6 = very dark brown). The points cluster together because the lineup members were deliberately chosen to match this description of the culprit. When the points cluster together, as they would if only these two features were considered, it is hard for the witness to discriminate the guilty suspect from the fillers. Therefore, additional feature dimensions must be considered.

Panel B: Representation of the same individuals when a third feature dimension (Eye Shape) is considered (0 = round to 6 = slanted). The Eye Color vs. Nose Size 2-D plot in panel A is now the floor of the 3-D plot in panel B (the points still cluster together on the floor), with the vertical axis representing the new dimension (Eye Shape). Because this feature was not included in the witness's description, the lineup members exhibit natural variability, so the points spread out along this dimension. Moreover, because eye shape is a feature of the culprit's face that the witness stored in memory but did not describe, now, only the guilty suspect is close to the culprit's representation in memory, generating a differentially strong memory-match signal (*m*). The critical point here is that to make an identification decision, the witness had to consider additional feature dimensions beyond those included in the description. Critically, considering additional feature dimensions individuates a face and is an example of elaborative processing. Elaborative processing makes memories incidentally (i.e., without intention to form a memory).

Elaborative processing creates a memory *incidentally* (i.e., without intention to form a memory). This is, in fact, the essence of the problem associated with testing a witness's memory for a suspect a second time. On that second test, a newly formed memory of the suspect will be accessible, even if the tested suspect is innocent, and the signal generated by the memory of the suspect's face might now be strong.

Source monitoring

The memory of a previously tested face is defined not only by its strength (i.e., by the magnitude of ***m***) but also by the memory of the context that accompanied the encoding of the face. Assigning context to the memory signal is known as *source attribution*, and it can be an error-prone process (Johnson, Hashtroudi, & Lindsay, 1993). That is, the witness might misattribute the strong memory signal to Source A when, in fact, the face was actually encountered in Source B. Testing memory for the first time using a police lineup almost seems tailor-made for inducing a source misattribution when memory is tested a second time.

An elaboratively processed face encoded during the initial test is not stored in a vacuum. Instead, it is encoded along with aspects of both the internal (i.e., reinstated) context and the external (i.e., testing) context (e.g., Nelson & Shiffrin, 2013; Cox & Shiffrin, 2017). These contexts are labeled Context A and Context B, respectively, in Figure 1. Critically, on the first test, only the culprit's face has been associated with the context of the crime (unlike any filler or any innocent suspect). However, on the second test, more faces will have been associated with that context, including the face of an innocent suspect being tested a second time. When Source A is again reinstated at the time of the second test ("Do you see the person who committed the crime?"), the activated content would now include not just the culprit's face but also the faces that were previously tested, including the face of the innocent suspect.

Typically, when the police conduct a second lineup with the same suspect and same witness, they use a new set of fillers. Therefore, in the typical case, the suspect will generate a *differentially* strong memory-match signal (potentially attributed to the wrong source) because only that face was tested previously. The differential familiarization of the suspect's face when memory is tested a second time violates the basic tenets of the "lineups-as-experiments" analogy (Wells & Luus, 1990). The idea is that when police investigators conduct a lineup, they are essentially performing an experiment to test their hypothesis that the suspect is guilty. As experimenters, they should adopt the same protocols that scientists adopt to ensure the integrity of their experiments. One of those protocols is to ensure that participants (witnesses in this analogy) are blind to the hypothesized outcome lest they perform in such a way as to please the experimenter. But if the suspect is the only person in common between the first and the second identification tests, then it is clear to the witness which person the police suspect of having committed the crime (namely the person in common to both procedures). Because the witness is no longer blind to the suspect's identity on the second test, the lineup is inherently biased against the suspect.

Empirical Studies of Testing Memory a Second Time

In light of the foregoing theoretical considerations, the memory signal generated by the innocent suspect's face will likely be stronger on a second test involving the same witness as a result of the witness having observed the suspect on the first lineup test. The relevant empirical evidence unambiguously supports this theoretical prediction.

The memory-for-foils paradigm

An illuminating experimental design known as the *memory-for-foils* paradigm provides compelling experimental evidence that testing memory contaminates memory by leaving behind a trace of the tested items (Jacoby, Shimizu, Daniels, & Rhodes, 2005). In a typical recognition memory experiment in the basic science literature, participants are presented with a list of items to study (e.g., a list of words). On a later recognition memory test, those same items (now called "targets") are randomly intermixed with new items ("foils"), and each item is presented individually for a yes/no decision (i.e., "Did this item appear on the list, yes or no?"). Theoretically, the activated memory content against which each test item is compared consists of the items from the study list (Cox & Shiffrin, 2017). In a test like this, the targets are analogous to a guilty suspect because they were seen on the list and the foils are analogous to innocent suspects and fillers because they were not seen on the list.

After completing the recognition test, the participants are then unexpectedly asked to complete a second recognition memory test consisting of the foils from the first test randomly intermixed with a new set of foils. This time, they are instructed to say "yes" to the foils that appeared on the first test (those items are now the targets) and to say "no" to the new foils. Theoretically, the activated memory set against which test items are compared consists of the items from the just-completed recognition memory test (including the foils). Therefore, the foils, when tested on the surprise memory test, will generate a relatively strong memory-match signal.

Indeed, participants perform very well on that second unexpected test even though, when they first saw the foils (now targets), they were merely attempting to decide whether or not those items had appeared on a previous list, not attempting to memorize them. The foils were

elaboratively processed to answer the recognition memory question and were encoded incidentally.

This phenomenon is not limited to lists of words but occurs for faces tested in a lineup as well. In a study reported by Charman and Cahill (2012), participants first viewed a mock-crime video and were later tested using a standard 6-person simultaneous photo lineup. Still later, the participants were given a surprise memory test for the five fillers in the lineup. This final test was a list memory test consisting of 10 faces (the five fillers plus five new faces), with each face presented individually for a yes/no decision about whether it had been seen previously in the lineup. Keep in mind that during the lineup test, the participants were not attempting to memorize the faces; instead, they only made an identification decision about each face. On the final test, the results were striking: the hit rate (proportion of previously seen fillers correctly recognized as such) was 76%, whereas the false alarm rate (proportion of new faces incorrectly recognized as having been previously seen) was only 19%.[1]

It is worth briefly considering how these results are interpreted in terms of the standard signal detection model discussed earlier (Box 1) because it illustrates a key point about how memory contamination caused by testing memory should be conceptualized. The model holds that previously seen faces will generate a stronger memory signal, on average, than new faces. On the unexpected test, the memory signals generated by the new (previously unseen) fillers are conceptualized as having been drawn from a Gaussian distribution with a low mean (Fig. 3A). Because these memory signals are weak, on average, only a small percentage exceeds the witness's decision criterion by chance (19%). The memory signal generated by the previously

[1] The false alarm rate was not reported by Charman and Cahill (2012) because the focus of their analysis was different from ours, but the authors kindly provided us with the data.

seen fillers (fillers-to-targets) are conceptualized as having been drawn from a Gaussian distribution with a high mean (Fig. 3B). Because these memory signals are strong, on average, a much higher percentage (76%) exceeds the witness's decision criterion.

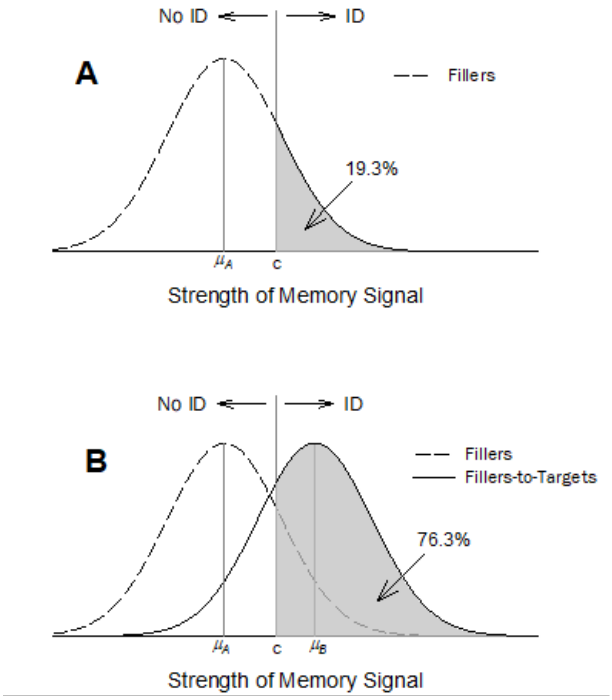


Figure 3. A. Standard signal detection interpretation of a false alarm rate of 19.3%. The memory signals generated by new fillers have a mean of μ_A, and the decision criterion (c) is placed well above that. B. The fillers-to-target distribution with mean μ_B has been added to the figure in panel A, showing the standard signal detection interpretation of a hit rate of 76.3%. Note that, relative to new foils, all the fillers-to-targets have had their memory strengths boosted.

For these data, $d' = 1.58$ (Box 1), which means that the participants could easily discriminate previously seen fillers from new fillers. Of most relevance to the issue under consideration here, this standard theoretical framework conceptualizes the memory distribution of the tested fillers (i.e., fillers-to-targets) as having been shifted upward relative to the fillers that had not yet been tested. In other words, it was not only the 76% of correctly recognized fillers-to-targets that had their memory signals strengthened by the initial test; the remaining 24% were strengthened (i.e., "contaminated") as well but not enough to exceed the decision

criterion. Thus, the face was rejected on a second test, but perhaps with less confidence than would otherwise have been the case. The take-home message is that, theoretically, testing memory contaminated memory for *all* the tested fillers.

Viewing mugshots

Brown, Deffenbacher & Sturgill (1977) had participants observe two separate groups of five strangers (henceforth “criminals”). One and a half hours later, they viewed 15 mugshots, including 5 people who were criminals and others who were seen for the first time. A week later, the participants were presented with lineups (a second memory test) and asked to identify the criminals from the initial in-person encounter. The experience of viewing the mugshots had a clear effect on memory. On the lineup test, of people who had never been seen before at all, the rate of mistaken identification was 8%. However, if a lineup member’s mugshot had been seen in the interim (but not during the original experience), the chances of being falsely identified as a criminal rose to 20%. Thus, the strong memory signals associated with the misidentified mugshot-only faces were misattributed as having been caused by the original experience involving the criminals (an example of source misattribution).

The findings reported by Brown et al. (1977) were reinforced by a more recent study reported by Goodsell, Gronlund, & Neuschatz (2015). Participants watched a short video clip of someone entering an office, after which they were randomly assigned to the mug shot condition or to the no-mug shot control condition. Those assigned to the mug shot condition viewed 50 mug shots of people matched to the description of the culprit from the video, and they were asked to search for the perpetrator. All participants returned after a 48-hour delay and viewed either a target present (TP) lineup or a target absent (TA) lineup.

If a previously seen mugshot photo in a lineup was one that the participant had previously picked, then that photo was (a) identified as the perpetrator from TP lineups much more often that the actual perpetrator (.70 vs. .08) and (b) identified from TA lineups with a very high probability (.81). If the previously seen mugshot photo was *not* the one that the participant had picked, then that photo was (a) *still* identified as the perpetrator from TP lineups more often that the actual perpetrator (.28 vs. .18) and (b) identified from TA lineups with a high probability of .38 (more than double the false suspect ID rate from the control condition). Thus, memory was contaminated by the initial mugshot test whether or not the mugshot face appearing in the later lineup had been previously identified (see related findings reported by Memon, Hope, Barrett, & Bull, 2002).

Testing a suspect a second time

Conceptually similar effects are observed when the initial test consists of viewing a lineup rather than mugshots. In Steblay, Tix, & Benson (2013), participants viewed a video crime and then attempted to identify the culprit from two 6-person lineups separated by a 2-week retention interval. The suspect (guilty or innocent) was common to both lineups. In the absence of contamination from the first test, the expectation would be that the guilty suspect ID rate would decline substantially after two weeks (due to forgetting) and the false ID rate would remain largely unchanged or increase slightly. For example, with similar retention intervals using a between-subjects design, Palmer et al. (2003) found the correct ID rate dropped from .60 to .51 (immediate to delayed), $p = .052$, whereas a slight increase in the false ID rate did not approach significance. By contrast, Steblay at el. (2013) found that when witnesses were tested both immediately and after a delay, the guilty suspect ID rate *increased* on the delayed test, albeit non-significantly (instead of exhibiting the expected decrease due to forgetting), and the false ID

rate now increased substantially from .21 to .31, $p = .03$. Thus, having seen the suspect in an earlier lineup contaminated memory, placing both innocent and guilty suspects at greater risk of being identified on a second test than would otherwise be the case.

Testing a witness's memory for a suspect a second time might not be problematic if, on the second test, not only was the same suspect included in the lineup but also the same fillers. In that case, everyone's face would generate an elevated memory signal compared to the first test, and no one would stand out. Lin, Strube, and Roediger (2019) conducted this very experiment and found that, even then, nothing was gained by conducting the second test. Instead, witnesses simply became more willing to choose but without improving accuracy.

This issue is specific to forensic memory evidence

In the forensic context, the problem associated with repeated testing is specific to *memory* evidence. For example, repeatedly comparing latent fingerprints lifted from a crime scene to the known fingerprints of a suspect is not problematic and can even serve the cause of justice (e.g., fingerprint examiners can double-check their work) because the test itself does not change the evidence. By contrast, repeated tests of memory are unlikely to serve the cause of justice because testing changes memory (Wells et al., 2020). If comparing latent prints and known prints from a suspect altered the latent prints in such a way as to more closely resemble the fingerprints of the suspect, it seems reasonable to suppose that any fingerprint test after the first would be viewed with suspicion and perhaps excluded from consideration. Although this kind of contamination does not occur with fingerprints, it does occur with "face prints" (the memory of the culprit in the brain of the eyewitness).

Of course, as noted earlier, such contamination can occur even before the first official memory test conducted by the police, so it makes sense to take steps to prevent that from happening. In this regard, Recommendation #1 from Wells et al. (2020) is relevant. The recommendation is to conduct a prelinup interview with the witness in which the witness is instructed to avoid attempting to identify the culprit on their own. If the witness has already done so and has encountered the suspect's photo (e.g., on social media), thereby contaminating memory prior to the first official test, it is also important to document that fact.

Memory contamination is not the only problem

By focusing on memory contamination resulting from the initial test of memory, we do not mean to imply that it is the only problem associated with testing memory more than once. Far from it. For example, as much prior research has shown, the risk to an innocent suspect associated with multiple testing is greatly compounded when suggestive procedures are used and/or when feedback to the witness is provided (Wells & Bradfield, 1998). If, for example, the witness misidentifies the innocent suspect with low confidence from a fair lineup, subsequent feedback from the police can quickly convert it to high confidence (e.g., if the police say "good job, we were pretty sure it was him"). In addition, other memory-contaminating events, such as seeing the face of the suspect again in pre-trial hearings or in news stories will further strengthen the memory signal generated by the defendant's face at trial. In addition, if the witness discusses independent evidence against the suspect with prosecutors, it will help to cement the source misattribution according to which the strong memory signal reflects having originally seen the suspect commit the crime.

All of this would be avoided by testing memory only once, thereby strengthening the rationale for the new test-memory-once recommendation in Wells et al. (2020). The new point

we are emphasizing here—one that has not received enough attention in the past—is that the witness's memory is *already contaminated* as a result of having taken the first test, even if pristine procedures were followed and even if none of the just-described additional factors exacerbated the problem (as difficult as that might be to imagine).

On the First Test, Confidence Protects Innocent Suspects

On the first (uncontaminated) test using a proper lineup, confidence is more likely to protect than imperil innocent suspects. As noted earlier (Box 1), signal detection theory predicts that decisions made with high confidence should be accurate, whereas decisions made with low confidence should be inaccurate. Related sequential sampling models (e.g., Pleskac & Busemeyer, 2010; Ratcliff, 1978; Ratcliff & Smith, 2004) make similar predictions about reaction time. That is, decisions made quickly should be accurate, whereas decisions made slowly should be less accurate. Empirically, these predictions have often been confirmed in list-memory studies conducted in the basic-science laboratory (e.g., Ratcliff & Murdock, 1976), in lineup studies conducted in the applied-science laboratory (Brewer, Caon, Todd, & Weber, 2006), and in lineup studies conducted in the real world (Seale-Carlisle, Colloff, Flowe, Wells, Wixted, & Mickes, 2019). For example, in a study involving actual eyewitnesses to a crime, Seale-Carlisle et al. (2019) reported that lineup decisions made rapidly (e.g., in 5 or 10 seconds) and with high confidence were estimated to be highly reliable, whereas decisions made slowly (e.g., 30 seconds or more) were much less reliable. This was true even in the rare case of a slow decision made with high confidence.

As unfortunate as a misidentification like this would be, keep in mind that the face of the innocent suspect does not actually correspond to the face stored in memory. Therefore, under optimal conditions, the strength of the memory-match signal, despite being randomly strong in a

particular case, is not likely to far exceed the witness's decision criterion. According to standard assumptions of signal detection theory (Box 1), and in accordance with much empirical evidence, under ideal testing conditions, misidentifications of the innocent (and of fillers) are usually made with something other than high confidence (Figure 4).

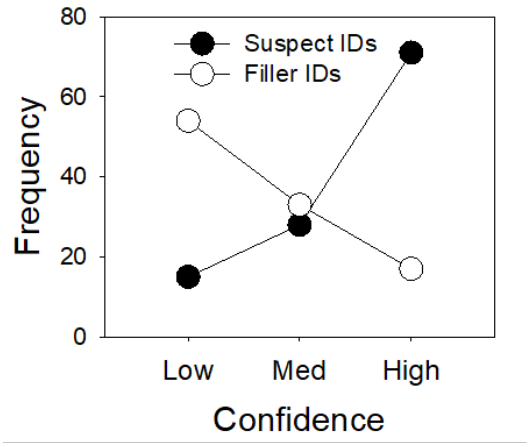


Figure 4. Number of Suspect IDs (filled circles) and Filler IDs (open circles) from 347 photo lineups administered to actual eyewitnesses in the Robbery Division of the Houston Police Department in 2013. The lineups were fair and were administered in double-blind fashion. Of interest here are the IDs made to known-innocents (i.e., the fillers), the large majority of which were made with low or medium confidence (from Figure 1B of Wixted et al, 2016).

Losing sight of the low confidence that might be associated with an initial ID (and losing sight of other red flags, such as initial filler identifications or lineup rejections) wastes an opportunity to protect innocent suspects. Identifications made with low confidence are known to be highly error prone, which means that a low-confidence identification should be regarded as an inconclusive test outcome (Wixted & Wells, 2017). This is why the police record of an identification made with low confidence should never be written as "the witness positively identified the suspect." Instead, the record should reflect the lack of confidence, and that lack of confidence should be taken to mean that the memory test was inconclusive. The phrase

"positively identified" is probably best reserved for cases where the witness is arguably *positive* that the identified individual is the culprit.

How should initial confidence be measured?

The best way to determine whether or not the witness was "positive" is an actively researched issue, and there is no consensus. Fortunately, the available research suggests that the different methods (e.g., a verbal scale, a 5-point numerical scale, a 100-point numerical scale, asking the witness to use their own words, etc.) may not matter very much. For example, Tekin & Roediger (2017) tested 4-, 5-, 20-, and 100-point scales and found that the different scales yielded similar (continuous) confidence-accuracy plots. In their words "The scales seem convertible from one to the other, and choice of scale range probably does not affect research into the relationship between confidence and accuracy" (p. 2).

Dodson & Dobolyi (2015) considered numerical vs. verbal confidence scales and concluded that "…confidence is calibrated with accuracy in a nearly identical manner when confidence is expressed with either a numeric scale or a verbal scale" (p. 267). In agreement with this claim, Tekin, Lin and Roediger (2018) recently compared 2- and 4-point verbal and numeric and found little difference between them. Very recently, Smalarz, Yang, and Wells (in press) and Mansour (2020) both asked participants to provide confidence in their own words or using a numerical scale. The results indicated that confidence was diagnostic of identification accuracy from a lineup either way, though Mansour (2020) found that verbal statements were more variable.

The upshot of the relevant research is that confidence should be assessed for an initial identification, as has been recommended for many years (Wells et al., 1998). Collecting a

confidence statement of some kind appears to be more important than exactly how it is done. Critically, without a confidence statement, it is not possible to know whether the initial ID was made with low confidence, in which case it is highly error prone. Because the initial test is the one that matters, it is essential to collect a confidence statement on that first test. Moreover, the entire identification procedure should be recorded on video (Recommendation #7 of Wells et al., 2020) so that all interested parties—the detective, prosecutor, judge, defense counsel, jury, and expert—can see and hear the confidence statement as it was captured in real time.

What if the First Test Involves a Bad Photo of the Suspect?

When a witness fails to identify the suspect or does so with uncertainty on the first test, the police sometimes conclude that the photo of the suspect was not a very good likeness to the face of the suspect. Therefore, they try again, conducting a second test using what they believe to be a better photo. Does a second test endanger an innocent suspect under these conditions?

If the “bad” photo is far from perfect but is nonetheless recognizably the suspect, and if the witness elaboratively processed that bad photo to make a memory-based decision about it (perhaps choosing not to identify that individual), then the witness has processed features that individuate the suspect’s face from other faces. The end product of such elaborative processing is an accessible memory record of that face. On the next memory test involving a better photo, those features will match the features that were encoded during the first test. This will have the effect of elevating the strength of the memory signal compared to what it otherwise would have been, thereby imperiling the innocent suspect. In other words, even a bad photo can contaminate memory if it is a recognizable photo of the suspect.

If the bad photo is instead not recognizable as the suspect, so much so that it might as well be a photo of a different person, then it is hard to see how memory contamination would occur. Therefore, in that case, a second test using a better photo would be reasonable. However, whether or not the photo is recognizable as the suspect is a judgment call. If no suspect ID is made, even a conscientious police investigator who strongly believes that the suspect is guilty might be inclined to honestly conclude that an imperfect-but-recognizable photo of the suspect was "bad," thereby justifying a second test. What can be done to protect against this alluring escape route from our main recommendation to test a witness's memory for a suspect only once?

The best solution would be to preserve the "bad" photo so that others can later judge for themselves whether or not it is a recognizable photo of the suspect. After all, this is not the only judgment call that an investigating officer has to make. The same officer will have judged the initial photo lineup to be fair, knowing that the photos would be preserved and later judged by others (e.g., by a jury at trial). Preserving the lineup photos incentivizes the investigating officer to exercise caution, ensuring that the lineup is fair. The same principle could be applied to the officer's judgment call about a photo of the suspect being so bad it might as well be a photo of another person. For example, at a pretrial hearing, if the court disagrees with that judgment call, then no later test involving the same suspect and eyewitness should be admissible as evidence.

Additionally, the issue can be tested empirically. For example, a sample of people can be presented with the first (allegedly "bad") photo of the suspect and then asked if they can pick that person out from the second lineup. If they cannot do so with greater-than-chance accuracy, then it would be reasonable to conclude that the photo was in fact bad enough that it did not taint the second identification procedure. However, if people can pick the suspect out of the second lineup with greater-than-chance accuracy, then the second lineup should be suppressed.

How Important was this Issue in the DNA Exoneration Cases?

Wixted and Wells (2017) argued that on an initial test of *uncontaminated* memory using a *pristine* lineup procedure, high confidence can imply high accuracy, and low confidence can imply low accuracy. Opinions differ as to the reliability of high-confidence IDs in the real world (e.g., Do those tests typically involve uncontaminated memory? Do they typically involve pristine procedures?), but the consensus view is that low-confidence IDs are highly error prone (i.e., at best, they are only weakly probative of guilt). This is true whether or not memory has already been contaminated by the time of the initial test and whether or not pristine procedures are used. Moreover, filler IDs and lineup rejections on the first test are not neutral outcomes but, if anything, are probative of innocence (Wells & Lindsay, 1980).

With that background in mind, consider an analysis reported by Garrett (2011) in his book *Convicting the Innocent*. As noted earlier, data from the Innocence Project show that eyewitness misidentifications contributed to ~70% of more than 375 wrongful convictions later overturned by DNA evidence. Garrett (2011) analyzed the trial records from 161 of those cases in which an eyewitness misidentified an innocent suspect. In the courtroom, at trial, the eyewitness identifications were almost all made with high confidence, which makes sense (otherwise, the prosecutor likely would not have put the witness on the stand). What did these witnesses do at the time of the initial identification? We do not have contemporaneous records, but what these eyewitness and police witnesses described at trial, according to Garrett's (2011) analysis yielded some interesting observations.

In 57% of the trial transcripts (92 of 161 cases), the witness who misidentified an innocent suspect with high confidence at trial recalled having initially done so with low confidence (34 cases), or they recalled having identified a filler, another suspect or no one at all

(64 cases), or they reported not having seen the culprit's face (15) cases (with some cases having more than one type of issue). We do not know what was said at these initial identifications, apart from what the witnesses later recounted at trial. However, to the extent that their recollections are accurate, these initial identifications were highly problematic not only because the suspect was not confidently identified but for other reasons as well (many of these lineups also involved highly suggestive procedures). This is a problem because IDs made with low confidence are known to be highly error prone. As Garrett (2011) put it, this can provide "a glaring sign that the identification was not reliable" (p. 64). Low-confidence IDs, as well as non-identifications or filler identifications or identifications of other suspects provide an opportunity to protect an innocent yet ultimately misidentified suspect. Unfortunately, for the DNA exoneration cases involving an inconclusive outcome (or a contrary outcome) on the initial test, that opportunity was lost because the witness's memory was tested more than once.

How many of the remaining cases—the ones for which no testimony about the initial decision exists (43% of 161 cases)—also involved an initial outcome other than a high-confidence ID of the suspect? There is no way to know, in part because in only four cases was the procedure recorded; at a minimum, the evidence reported by Garrett (2011) is consistent with the idea that a sizable fraction of consequential eyewitness misidentifications began with something other than a conclusive (i.e., high-confidence) identification of the suspect. Indeed, as illustrated earlier in Figure 1, it is also consistent with the findings of a police department field study in which misidentifications of known-innocents (fillers) were much more likely to have been made with low or medium confidence than high confidence. We turn now to three cases that illustrate how important this issue is.

Three Illustrative Cases

John Jerome White

On August 11, 1979, a man broke into a Manchester, Georgia, house and raped a 74-year-old woman asleep on her couch (details about this case are available here and here). Based on the description of the culprit provided by the victim, the police created a composite sketch. A Georgia Bureau of Investigation agent happened to be investigating a 19-year-old man named Jerome White on another charge, and he thought the sketch resembled White. A week later, the victim picked White out of a photo array, but she was not completely certain (saying she was "almost positive" he was the attacker).

Perhaps because of those initial signs of uncertainty, she was later administered a live lineup (Figure 5). White was the only person to appear in both the photo lineup and live lineup (none of the fillers were repeated), so his face had been differentially familiarized as a result of the initial photo lineup test. Based on all theoretical and empirical considerations we have reviewed to this point, it would not be surprising to learn that the victim identified White again from this lineup, and she did. What makes this case remarkable, however, is who one of the fillers in the lineup turned out to be.

Figure 5. Live lineup administered to the victim following an initial photo lineup in which she identified Jerome White. White is in the middle, and the actual rapist (James Parham) is the man on the right.

The victim originally told the police that her attacker was “well built” and had a “round face,” a description that does not apply to Jerome White (Figure 5). However, it does apply to one of the fillers in the lineup, namely, the man at the far right in Figure 5. He was not a suspect, but he happened to be in jail at the time, so he was selected to fill out the lineup. Incredibly, many years later, DNA evidence indicated that he was the one who actually committed the rape. Yet after seeing White’s face in the initial photo lineup, choosing him, and (evidently) making a source misattribution, he was now the face that came to mind when the victim was asked if she sees the man who raped her in the lineup.

At his trial in 1980, the victim conclusively identified White as the man who had raped her (“that’s him”). And why not? She had now seen his face on multiple previous tests, and the strong sense of familiarity was, in her mind, sourced to the initial crime (not to the lineup tests). The police and prosecutors presumably also reinforced her choice, not to intentionally create an

injustice, but to reassure her. Unfortunately, such reassurance only serves to inflate confidence (Wells & Bradfield, 1998). Any doubts the witness had at the initial lineup vanished by the time of the trial. White spent more than 22 years in prison before finally being exonerated by DNA evidence in 2007. Based on the same DNA evidence that exonerated White, prosecutors charged James Parham (the man on the far right in Figure 5) with the rape. He pled guilty and was sentenced to 20 years in prison.

Steven Gary Titus

On October 12, 1980, Seattle police received a report that a man had raped a female hitchhiker (details about this case can be found here and a Ted talk about it can be found here). Steve Titus was a restaurant manager in Seattle at the time, and he was on his way home from a restaurant with his fiancé when his car was stopped by a police officer because it resembled the car that was driven by the rapist. Titus also fit the description of the rapist provided by the victim. When later presented with a photo lineup, the victim identified Titus as her attacker, stating “This one is the closest one.” It might very well be the case that Titus provided the closest match to her memory of the culprit (thereby generating the strongest memory signal of the faces in the lineup), but her wording is indicative of low confidence, not high confidence. Yet when Steve Titus was put on trial for rape, the witness’s uncertainty had vanished. When she got on the witness stand, she identified Titus with high confidence. By then, not only did the witness have a memory of Titus based (at a minimum) on the initial lineup test, likely ensuring a stronger memory signal the next time his face was seen, but she may also have been informed of other reasons why police and prosecutors thought he was guilty (further inflating confidence), and the courtroom identification test itself is inherently suggestive (inflating confidence still further). Based largely on that confident testimony, Titus was found guilty.

According to an article in the New York Times, a few months after Titus was convicted, new evidence suggested that a different suspect was responsible for a series of rapes in the area. When the rape victim saw the photograph of the new suspect, she realized that he was the one who had actually raped her. At that point, she began to cry and said “Oh my God, what have I done to Mr. Titus?” However, the key mistake was made by other actors in the criminal justice system, not the witness, because they tested memory for the suspect (Titus) more than once, ignoring her initial low-confidence ID. Instead of relying on the first test only, they unwittingly relied on contaminated memory evidence at trial to win what turned out to be a wrongful conviction.

Fortunately, Titus avoided a long stint in prison, but the story does not otherwise have a happy ending. Embittered by his wrongful conviction and the financial ruin it caused (including large legal fees and the loss of his job), he decided to file a lawsuit against the Port of Seattle police. Sadly, just before that case was to be heard, Titus died of heart failure at the age of 35.

Charles Don Flores

On the morning of January 29, 1998, witness Jill Barganier saw two people get out of a car outside the home of her neighbor, Elizabeth Black, who was murdered shortly thereafter (details about this case can be found in legal documents posted here). Bargainer described both as white males with long, shoulder-length hair. Another neighbor independently described seeing two white males get out of the car and enter the Black’s house that morning. When presented with an initial photo lineup containing the main police suspect, a man named Richard Lynn Childs (a white man with long hair down to his shoulders), Barganier immediately identified him with high confidence. Childs owned a handgun of the same caliber used to murder Black, and he owned a conspicuously painted car that multiple eyewitnesses saw parked near Black’s home the

morning of the murder. He also eventually signed a confession admitting that he shot Black. This was an initial identification made quickly and with high confidence, and all indications are that it was a reliable ID.

Who was the other man Bargainer saw getting out the car that morning? The police suspected Charles Don Flores because he was a known associate of Childs and had been engaged in a drug deal with him in the hours before the murder. Flores was a heavyset Hispanic man with a crew cut and therefore did not match the description of the accomplice provided by the witness. Nevertheless, the police placed his photo in a lineup with other Hispanic men as fillers and presented it to the witness. Quite understandably, the witness did not identify anyone (i.e., she rejected the lineup). This makes sense because it is hard to see why photos of large Hispanic males with short hair would generate a strong memory-match signal when compared against the memory of a white male with long hair stored in the witness's brain. Thus, on the initial test, her failure to identify Flores provides no evidence of guilt and, if anything, provides evidence of innocence.

Nevertheless, at the trial, Jill Bargainer was certain that Flores was the man she saw that morning with Childs. Multiple factors presumably contributed to her high confidence, beginning with the memory-based decision she made about Flores on the first test (from that moment on, she likely had a representation of his face in her memory), perhaps continuing with news stories in which his face was shown, and culminating in the suggestive memory test performed at trial. For all these reasons, the only relevant eyewitness evidence was her rejection of the initial lineup.

In addition to the "direct" courtroom evidence provided by the eyewitness, there is indirect evidence against Flores as well. For example, in the days following the crime, he torched

the conspicuous paint job on the car driven by Childs the morning of the murder (presumably to make it harder for the police to find), and he fled to Mexico when he learned that the police were looking for him (i.e., he "acted guilty"). This information, if Bargainer were aware of it, would have also served to bolster her confidence by the time of the trial. Childs did not testify about his accomplice at the time of the trial and has not done so to this day.[2]

Despite some independent evidence of guilt, by all accounts, it was the testimony of an extremely credible and highly confident eyewitness that led to the conviction of Flores. In Texas, murder is a capital crime, and an accomplice to a murder is as guilty as the triggerman (Childs). Therefore, Flores was sentenced to death. He has been on death row for over 21 years, and his appeal to the U.S. Supreme Court was denied on January 22, 2021. What may be his final appeal recently filed in the Texas Court of Criminal Appeals.

The most remarkable fact about this case is that the eyewitness evidence that is mainly responsible for sending him to death row (namely, the witness's confident testimony at trial) is actually probative of innocence when properly understood (i.e., her initial description of the accomplice and her rejection of the initial lineup). In this case, police and prosecutors obviously failed to appreciate that only the first test counts.

A Simple Reform: Test a Witness's Memory for a Suspect Only Once

Presenting the face of a stranger on an eyewitness identification test contaminates the witness's memory for that individual. Such contamination is difficult to avoid, and if it occurs, there is no way to undo it (i.e., there is no way to decontaminate memory). If the witness's memory for that individual suspect is tested again, the suspect's face will generate a stronger memory signal than it otherwise would. As noted earlier, the fact that memory has been

[2] In a plea bargain, Childs was sentenced to 30 years in prison and was released after serving 16 years (i.e., he is a free man today).

contaminated does not necessarily mean that the contaminated memory signal will be strong enough to exceed the criterion for making an identification. However, even in that case, memory has been irretrievably contaminated. On any later test, due to source misattribution, witnesses are at risk of responding to the elevated memory signal as if it were based on a memory formed at the time of the crime. By the time of trial, a variety of factors over and above the contaminating effects of testing memory more than once (feedback from the police, seeing the suspects face on the news, etc.) will have likely exacerbated the problem.

In contrast to this science-based theoretical perspective, judges often have a different legal perspective. As noted by Garrett (2012), they often embrace the catastrophically mistaken idea that, following the initial test, it is possible to conduct an "independent" test of memory, as if testing the match between a suspect's face and the witness's memory of the culprit multiple times is like testing the match between a suspect's fingerprints and the latent fingerprints lifted from a crime scene multiple times. However, as noted earlier, fingerprints do not change from the first test to the second; memories do. Therefore, once it has been tested and contaminated, it is not possible to perform a second independent test of the memory of a stranger's face that was formed during the commission of the crime.

The only barrier to implementing this recommended reform (test a witness's memory for a suspect only once) is a faulty theory in the minds of various actors in the criminal justice system. It therefore follows that implementing this reform should be much simpler than implementing other reforms that require training officers to administer eyewitness identification tests properly. To implement this newly proposed reform, the only training that is required is for policymakers to change their thinking about how memory works and to understand that testing memory for a suspect carries the high risk of irretrievably contaminating memory of that suspect.

Considering how many wrongful convictions based on eyewitness misidentification might have been avoided by understanding this simple idea—and considering how many might be avoided going forward—implementing this reform should be an urgent priority.

A Final Word about Courtroom IDs

Because testing memory for a suspect is likely to contaminate memory for that face, a memory test conducted in the courtroom is likely to be a test of contaminated memory, by which time many additional factors exacerbate the problem. There may be rare exceptions (e.g., when the first test of memory for the defendant occurs from the witness stand, at trial, or when the prior test involved a photo that is not recognizable as the defendant). However, even in cases like that, despite avoiding the problem of memory contamination, a courtroom ID would still be problematic due to its inherently suggestive nature (Wells & Luus, 1990). It is inherently suggestive because only one person is sitting next to the defense attorney, making it plainly obvious to all that prosecutors believe they have enough evidence to be convinced that this is the person who committed the crime.

When it comes to eyewitness identifications, the courts often have it exactly backwards, sometimes excluding earlier tests (including the all-important initial test) while allowing in court IDs based on memory that (unbeknownst to the judge) has likely been contaminated by events that occurred after the crime. As Garrett (2012) put it: “Today courts almost always allow courtroom identifications, but they sometimes bar prior identifications. Instead, courts should per se exclude courtroom identifications if there was a prior identification, but they should sometimes admit out-of-court identifications” (p. 457). Perhaps exceptions could be made for rare circumstances like those mentioned above, but the point is that it makes sense for courts to exclude forensic evidence that has likely been contaminated in a way that is prejudicial to the

defendant instead of making an exception for contaminated eyewitness evidence by routinely allowing it.

In addition to excluding courtroom identifications, except under presumably rare circumstances, the only out-of-court identification that should be admitted is the *first* one. Only the first test should be admitted for the same reason the court might exclude other kinds of forensic evidence that had likely been contaminated. Barring unusual circumstances (e.g., the witness did even look at the suspect on the first lineup test, or the photo used in the first lineup test was not even recognizable as the suspect), that first test provides the only relevant memory evidence. Even the first official test may involve contaminated memory (e.g., if the witness found a photo of the suspect on social memory before viewing the photo lineup), but the first test unarguably provides the best chance to test uncontaminated memory. This simple reform, had it been implemented long ago, could have prevented many (perhaps most) of the wrongful convictions that occurred not because of eyewitness misidentification but because memory was tested more than once.

References

Abudarham, N., Shkillera, L., & Yovel, G. (2019) Critical features for face recognition. *Cognition 182*, 73-83.

Brewer, N., Caon, A., Todd, C., & Weber, N. (2006). Eyewitness Identification Accuracy and Response Latency. *Law and Human Behavior, 30*, 31–50.

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied, 12*, 11–30.

Brown, E., Deffenbacher, K., & Sturgill, W. (1977) Memory for faces and the circumstances of encounter. *Journal of Applied Psychology, 62*, 311-318.

Charman, S. D., & Cahill, B. S. (2012). Witnesses' memories for lineup fillers postdicts their identification accuracy. *Journal of Applied Research in Memory and Cognition, 1*, 11–17.

Chua, K.-W., Richler, J.J., & Gauthier, I. (2015). Holistic processing from learned attention to parts. *Journal of Experimental Psychology: General, 144*, 723–729.

Cox, G. E., & Shiffrin, R. M. (2017). A dynamic approach to recognition memory. *Psychological Review, 124*, 795–860.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, 268–294.

Davis, D. & Loftus, E. F. (2018). Eyewitness science in the 21st century: What do we know and where do we go from here? In J. T. Wixted (Series Ed.) & E. A. Phelps and L. Davachi (Vol. Eds.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (*4th Ed, Vol. 1*): *Learning & Memory*. Hoboken, NJ: John Wiley & Sons, Inc.

Dodson, C. S. & Dobolyi, D. G. (2015). Misinterpreting eyewitness expressions of confidence: The featural justification effect. *Law and Human Behavior, 39*, 266–280.

Fechner, G. (1860/1966). *Elements of psychophysics. Vol. 1*. Holt, Rinehart and Winston: New York.

Fisher, R. P., & Geiselman, R. E. (1992). *Memory-enhancing techniques for investigative interviewing: The cognitive interview*. Springfield, IL: Charles C. Thomas.

Garrett, B. L. (2011). *Convicting the Innocent: Where Criminal Prosecutions Go Wrong*. Cambridge, MA: Harvard University Press.

Garrett, B. L. (2012). Eyewitnesses and exclusion. *Vanderbilt Law Review, 65*, 451-506.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*, 1–67.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology, 66*, 325–331.

Goodsell, C. A., Gronlund, S. D., & Neuschatz, J. S. (2015). Investigating mug shot commitment. *Psychology, Crime & Law, 21*, 219–233.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* John Wiley.

Innocence Project (2020). Understand the causes: the causes of wrongful conviction. New York:

Innocence Project. https://www.innocenceproject.org/eyewitness-identification-reform/. Accessed November 11, 2020.

Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review, 12*, 852–857.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*, 3–28.

Kellen, D., Winiger, S., Dunn, J., Singmann, H. (2021). Testing the Foundations of Signal Detection Theory in Recognition Memory. *Psychological Review* (In Press).

Lin, W., Strube, M. J., & Roediger, H. L. (2019). The effects of repeated lineups and delay on eyewitness identification. *Cognitive research: principles and implications, 4*, 1.

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 19–31.

Loftus, E. F. & Palmer, J. C. (1974) Reconstruction of automobile destruction: An example of the interaction between language and memory. J*ournal of Verbal Learning and Verbal Behavior, 13*, 585 -589.

Loftus, E. F. & Pickrell, J. E. (1995). The formation of false memories. *Psychiatric Annals, 25*, 720–725.

Mansour, J. (2020). The confidence-accuracy relationship using scale versus other methods of assessing confidence. *Journal of Applied Research in Memory and Cognition, 9*, 215-

231.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*, 724–760.

McKone, E., & Yovel, G. (2009). Why does picture-plane inversion sometimes dissociate perception of features and spacing in faces, and sometimes not? Toward a new theory of holistic processing. *Psychonomic Bulletin Review, 16*, 778–797.

Memon, A., Hope, L., Bartlett, J., & Bull, R. (2002). Eyewitness recognition errors: The effects of mugshot viewing and choosing in young and old adults. *Memory & Cognition, 30*, 1219–1227.

Morgan, C. A., Southwick, S., Steffian, G., Hazlett, G. A., & Loftus, E. F. (2013). Misinformation can influence memory for recently experienced, highly stressful events. *International Journal of Law and Psychiatry, 36*, 11-17.

Munsterberg, H. (1908). *On the Witness Stand*. McClure: New York.

National Research Council (2014). *Identifying the culprit: Assessing eyewitness identification*. Washington, DC: The National Academies Press.

Nelson, A. B., & Shiffrin, R. M. (2013). The co-evolution of knowledge and event memory. *Psychological Review, 120*, 356–394.

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied, 19*, 55-71.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117*, 864–901.

Police Executive Research Forum (2013). A National Survey of Eyewitness Identification Procedures in Law Enforcement Agencies. Retrieved March 29, 2016, from http://www.policeforum.org/

Porter, S., Yuille, J. C., & Lehman, D. R. (1999). The nature of real, implanted, and fabricated memories for emotional childhood events: Implications for the recovered memory debate. *Law and Human Behavior, 23*(5), 517–537.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59–108.

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review, 83*, 190–214.

Ratcliff, R. and Smith, P.L. (2004) A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111*, 333–367

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long term-retention. *Psychological Science, 17*, 249–255.

Seale-Carlisle, T. M., Colloff, M. F., Flowe, H. D., Wells, W., Wixted, J. T. & Mickes, L. (2019). Confidence and response time as indicators of eyewitness identification accuracy in the lab and in the real world. *Journal of Applied Research in Memory and Cognition, 8*, 420-428.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM – retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*, 145–166.

Smalarz, L., Yang, Y., & Wells, G. L. (in press). Words vs. numbers: A comparison of the diagnostic utility of eyewitnesses' verbal and numeric confidence statements. *Law and Human Behavior*.

Steblay, N. K., & Dysart, J. E. (2016). Repeated eyewitness identification procedures with the same suspect. *Journal of Applied Research in Memory and Cognition, 5*, 284-289.

Steblay, N. K., Tix, R. W., & Benson, S. L. (2013). Double exposure: The effects of repeated identification lineups on eyewitness accuracy. *Applied Cognitive Psychology, 27*, 644–654

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology, 46A*, 225-245.

Tanaka, J. W., & Simonyi, D. (2016). The "parts and wholes" of face recognition: a review of the literature. *Quarterly Journal of Experimental Psychology, 69*, 1876–1889.

Technical Working Group for Eyewitness Evidence (1999). Eyewitness evidence: A guide for law enforcement (Report No. NCJ 178240). Washington, DC: United States Department of Justice, Offices of Justice Programs. http://www.ncjrs.gov/pdffiles1/nij/178240.pdf

Tekin, E., & Roediger, H. L. (2017). The range of confidence scales does not affect the relationship between confidence and accuracy in recognition memory. *Cognitive Research: Principles and Implications, 2*, 49.

Tekin, E., Lin, W., & Roediger, H. L. (2018). he relationship between confidence and accuracy with verbal and verbal + numeric confidence scales. *Cognitive Research: Principles and Implications, 3:41*. https://doi.org/10.1186/s41235-018-0134-3

Tulving, Endel (1983). *Elements of episodic memory*. Oxford: Clarendon Press.

Tulving, E. & Thomson, D. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*, 352–373.

Tupper, N., Sauerland, M., Sauer, J. D., & Hope, L. (2019). Eyewitness identification procedures for multiple perpetrator crimes: A survey of police in Sweden, Belgium, and the Netherlands. *Psychology, Crime & Law*, DOI: 10.1080/1068316X.2019.1611828

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology*, *43*, 161-204.

Valentine, T., Lewis, M. B., & Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *The Quarterly Journal of Experimental Psychology*, *69*, 1996-2019.

Watkins, M. J., Ho, E., & Tulving, E. (1976). Context effects in recognition memory for faces. *Journal of Verbal Learning & Verbal Behavior, 15*, 505–517.

Watkins, M. J., & Watkins, O. C. (1976). Cue-overload theory and the method of interpolated attributes. *Bulletin of the Psychonomic Society, 7*, 289–291.

Wells, G. L. & Bradfield, A. L. (1998). “Good, you identified the suspect:” Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83*, 360-376.

Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A., & Wixted, J. T. (2020). Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law and Human Behavior, 44*, 3-36.

Wells, G. L., & Lindsay, R. C. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin, 88*(3), 776–784.

Wells, G. L., & Luus, C. E. (1990). Police lineups as experiments: Social methodology as a framework for properly conducted lineups. *Personality and Social Psychology Bulletin, 16*, 106–117.

Wells, G. L., Small, M., Penrod, S. J., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior, 22*, 603–647.+` 2

Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*, 201-233.

Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E. & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences, 113*, 304-309.

Wixted, J. T., Vul, E., Mickes, L. & Wilson, B. W. (2018). Models of lineup memory. *Cognitive Psychology, 105*, 81-114.

Wixted, J. T. & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*, 10-65.